Speech Acoustic Modelling From Raw Phase Spectrum

SPEECH ACOUSTIC MODELLING FROM RAW PHASE SPECTRUM

Erfan Loweimi [1]*, Zoran Cvetkovic* [2]*, Peter Bell* [1] *and Steve Renals* [1]

[1] Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] Department of Engineering, King's College London, UK

ABSTRACT

Magnitude spectrum-based features are the most widely employed front-ends for acoustic modelling in automatic speech recognition (ASR) systems. In this paper, we investigate the possibility and efficacy of acoustic modelling using the raw short-time phase spectrum. In particular, we study the usefulness of the raw wrapped, unwrapped and minimum-phase phase spectra as well as the phase of the source and filter components for acoustic modelling. Furthermore, we explore the effectiveness of simultaneous deployment of the vocal tract and excitation components of the raw phase spectrum using multi-head CNNs and investigate multiple information fusion schemes. This paves the way for developing an effective phase-based multi-stream information processing systems for speech recognition. The performance, even for wrapped phase with a noise-like shape, is comparable to or better than the magnitude-based classic features, and up to 4.8% WER has been achieved in the WSJ (Eval-92) task.

Index Terms: Raw phase spectrum, phase-based source-filter separation, multi-head CNNs, acoustic modelling, ASR

1. INTRODUCTION

The phase spectrum is not an appealing part of the Fourier transform (FT) for processing the speech signal. Owing to the *phase wrapping* phenomenon, this spectrum has a complicated shape, difficult to be understood, modelled and linked to the known properties of the human speech production and perception systems. In addition, it may be argued that the phase spectrum is perceptually unimportant, primarily due to some historical bias originating from Ohm's phase law [1]. Further, it is demonstrated that the phase-only reconstructed speech stimuli are readily intelligible only when signal is decomposed into long frames (e.g. 512ms in [2]) while because of non-stationarity, speech is analysed on a short-term basis.

Nevertheless, among others, the phase spectrum has been applied in speech analysis and feature extraction. *Group delay* (GD), the negative derivative of the unwrapped phase, is a widely used representation of this spectrum. Under some controlled conditions, the GD resembles the magnitude spectrum and renders a higher spectral resolution. Such similarity allows for substituting the magnitude spectrum with the GD in the magnitude-based algorithms, e.g. in the MFCC pipeline, which helps towards extracting phase-based features.

For speech signals, however, the vanilla GD is too spiky and should be somehow smoothed to become practically useful. Dealing with its spikiness and turning the GD into features using algorithms inspired by the magnitude-based workflows, have been widely explored in the literature, e.g. [3–9].

This paper aims to employ the raw short-time phase spectrum for ASR. The novelties are three fold: first, the possibility and efficacy of acoustic modelling using the raw phase spectrum is investigated. Such an approach bypasses feature engineering process which inextricably leads to task-useful information loss. As a matter of fact, the complexity of the phase structure and lack of insight into how it encodes information, further complicates crafting an effective feature extraction pipeline. Raw phase modelling not only contributes towards task-specific information filtering but also paves the way for representing phase information in an optimal format.

Second, we combine the phase's source and filter components using multi-head CNNs and explore several fusion schemes for building an effective phase-based multi-stream information processing system. This allows for processing each individual stream using a series of bespoke transformations and then fusing them at an optimal level of abstraction.

Third, we examine the usefulness of acoustic modelling from phase spectrum on a large vocabulary continuous speech recognition (LVCSR) task, namely WSJ [10]. The previous applications have been limited to small/medium range tasks.

Having reviewed the applications of the phase spectrum in ASR in Section 2, we briefly overview the source-filter separation in the phase domain in Section 3. Section 4 explores the "why" and "how" of recombining the phase spectra of the speech's vocal tract and excitation information streams using multi-head CNNs. Section 5 includes experimental results along with discussion and Section 6 concludes the paper.

2. PHASE SPECTRUM APPLICATIONS IN ASR

For the minimum-phase (MinPh) signals, the GD resembles the magnitude spectrum: local maximum at poles and minimum at zeros. Such resemblance enables replacing the magnitude with the GD in the workflow. Besides, some modifications might be required, too, e.g. in the MFCC pipeline,

Supported by EPSRC Project EP/R012180/1 (SpeechWave).

since the dynamic range of the GD is already limited, the log (or root compression) may be bypassed. However, the unprocessed GD of the speech signal is too spiky due to proximity of the zeros (associated with the excitation component) to the unit circle. This necessitates embedding some smoothing steps in the pipeline to make the GD practically applicable.

In [5], cepstral smoothing was utilised in computing the so-called *modified group delay* (MGD). MGD was initially proposed for speech analysis [5] and after further modifications employed in feature extraction for phone recognition [6]. In [7], chirp processing was employed for dealing with the spikiness of the GD, leading to *chirp group delay* (CGD). Its performance was evaluated on the Aurora-2 (A2) [11] (noisy connected-digit recognition) task.

Computing the GD after parametric signal modelling (primarily linear predictive coding (LPC)) is another possible solution for the spikiness. This approach was first explored in [3] for formant extraction. In [4] a relationship between the GD and cepstral coefficients was established. Then, a smoothed GD was computed after some cepstral weighting on top of the LPC ceptrum and the features were tested in a small isolated-word recognition task. In [8], GD of the LPC models was used for phase-based feature extraction and the corresponding features were evaluated on the A2 task.

Other phase-based representations such as its temporal derivative, namely *instantaneous frequency* (IF) [12, 13], and the so-called *product spectrum* (product of the GD and Periodogram) [14], have also been proposed and tested on digit recognition tasks. In [15], the ideas of noise compensation via vector Taylor series (VTS) [16] and generalised VTS [17] methods were extended to the product spectrum domain and successfully evaluated on the Aurora-4 (A4) [18] (medium vocabulary continuous speech recognition) task.

In [19], the statistical properties of the phase spectrum and phase-based representations were scrutinised. Moreover, usefulness of various types of statistical normalisation methods (histogram equalisation, Gaussianisation and mean-variance normalisation) of the phase-based features at different stages along the pipeline were investigated on the A2 task.

Source-filter separation in the phase domain [9, 20, 21], sheds further light on the structure of the speech's phase spectrum and set the stage for applying it in a wider range of applications. For example, in [22], the phase's source component was employed for fundamental frequency (F0) extraction. In [20], the filter component of the phase was utilised for feature extraction and tested on the A2 and A4 tasks.

The speech phase spectrum also has been implicitly employed for estimating the *complex ratio mask* [23] with applications in speech separation [23] and ASR [24].

3. PHASE-BASED SOURCE-FILTER SEPARATION

In this section, we briefly review source-filter separation in the phase domain [9, 20, 21]. Segregation of the excitation and vocal tract components provides two information streams

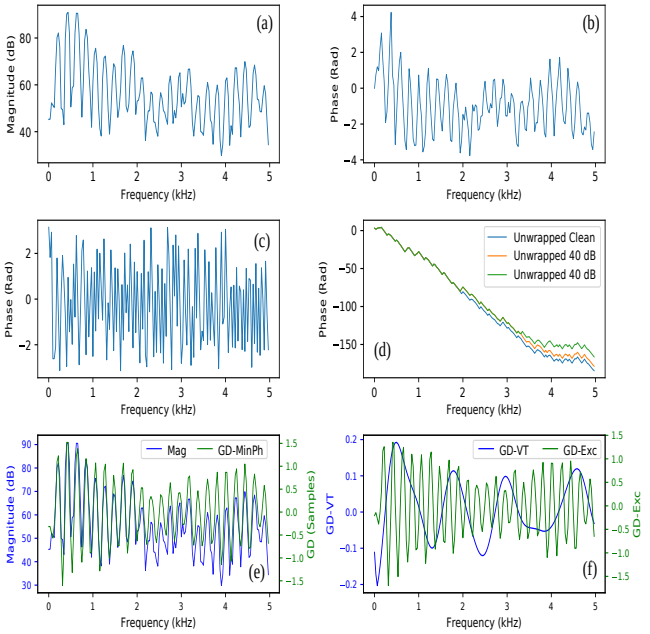

Fig. 1. Raw FT-based representations. (a) Magnitude spectrum, (b) phase of the MinPh component (c) wrapped phase, (d) unwrapped phase in three conditions, (e) GD of the MinPh component (green), (f) GD of filter (blue) and source (green).

which can be effectively processed via multi-head CNNs. The advantages of such recombination are discussed in Section 4.

To disentangle the speech's source and filter elements in the phase domain two questions should be addressed: How are these two components mixed in the phase domain? How can we manipulate the phase spectrum to separate them?

Assuming the speech production system is linear and the vocal tract (VT) and excitation (Exc) parts are independent (do not interact), for the frame $x[n]$ the following hold [9]

$$x[n] = x_{VT}[n] \ * \ x_{Exc}[n] \tag{1}$$

$$\log|X(\omega)| = \log|X_{VT}(\omega)| + \log|X_{Exc}(\omega)| \tag{2}$$

$$\arg\{X_{MinPh}(\omega)\} = -\frac{1}{2\pi}\log|X(\omega)| \ * \ \cot\left(\frac{\omega}{2}\right) \tag{3}$$

$$\arg\{X_{MinPh}(\omega)\} = \arg\{X_{VT}(\omega)\} + \arg\{X_{Exc}(\omega)\} \tag{4}$$

$$\mathrm{GD}_{MinPh}(\omega) = \mathrm{GD}_{VT}(\omega) + \mathrm{GD}_{Exc}(\omega) \tag{5}$$

where n, ω, X, $*$, cot, $|\cdot|$ and arg denote time, angular frequency, FT of $x[n]$, convolution operator, cotangent function, magnitude and unwrapped phase spectra, respectively.

As elaborated in [9,21], $\arg\{X_{MinPh}(\omega)\}$ (Fig. 1(b)) can be interpreted as a superposition of two components: a slowly varying part (*Trend*), modulating a rapidly oscillating *Fluctuation* element. The former is associated with the vocal tract and the latter pertains to the excitation part. Based on having different rates of change with respect to the independent variable (ω) and using the additivity in Eq. (4), the source and filter components can be separated by applying a proper low-pass filtering, as illustrated in Fig. 1(f).

Moreover, in [20] two modifications were suggested to improve the noise robustness: first, the log in the Hilbert transform (Eq. (3)) was replaced with the *generalised logarithmic function*. Second, the spectral derivative of the phase spectrum, i.e. the GD, was calculated via *regression filter* rather than the sample difference which is inherently noisy. The former was mainly helpful in feature extraction from the filter component for ASR [20] while the latter was instrumental in F0 extraction using the source component [22]. In this paper, we deploy both of these amendments.

4. SOURCE-FILTER RECOMBINATION THROUGH MULTI-HEAD CNNS

For acoustic modelling, the raw phase spectra of the source and filter components can be employed individually or simultaneously. The latter can take two forms: feeding a single-head CNN with their sum, namely the phase of the MinPh part (Eq. (4)) or using a two-head CNN and feeding one head with the filter and one head with the source component (Fig. 2).

Before dealing with the implementation aspects, let us discuss the intuition and advantages of employing a multi-head architecture. That is, what is the merit of separating the phase spectra of the source and filter components and then, recombining them using multi-head CNNs? Why do not just feed the CNN with their *sum* to take advantage of the information of both vocal tract and excitation components?

To answer these questions two points should be considered: first, the *importance* of the information content of each stream for a given task is different and, the optimal mixing weights are not digital (0/1) but fuzzy. Using the sum means giving the same weight (one) to each stream while it is a priori known that the weight of the VT part should be larger for an ASR task. In addition, as shown in Section 5, although the source component unsurprisingly results in poorer WERs than the VT element in ASR, its performance is still much better than random choice. This implies that although it is not as important as the VT part, it still includes some potentially beneficial information. Therefore, its optimal weight should be larger than zero. Using the sum of the source and filter as input, could restrict the effective handling of this issue.

Second, regardless of the information importance, the information generation processes underlying each stream, encode differing *types* of information in various *formats* and arrangements. Optimal handling of such variability involves using different bespoke sequence of transforms for each stream to extract abstract representations, ideally containing only task-useful information, devoid of task-irrelevant one. Multi-head CNNs can cope well with this challenge, too.

Having processed each information stream individually, how they should be fused? As illustrated in Fig. 2, we employ a multi-stream information processing system composed of a cascade of convolutional and fully-connected (FC) layers. Information fusion is carried out by concatenating the streams

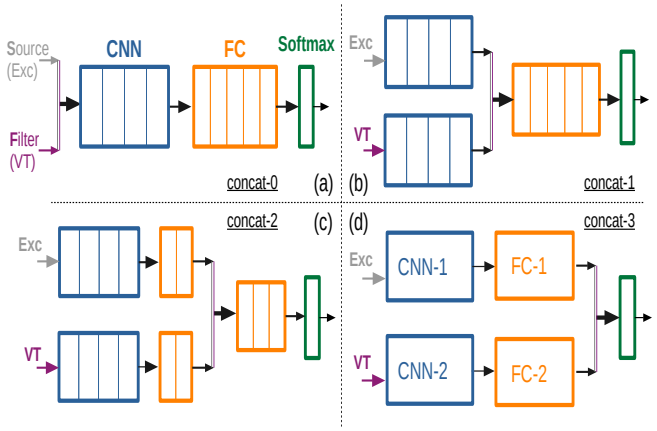

Fig. 2. Multi-head acoustic modelling using a cascade of convolutional and FC layers. Fusion (concatenation) of the raw phase spectra of the source and filter components at different levels: (a) concat-0, (b) concat-1, (c) concat-2, (d) concat-3.

at four plausible levels: the input level (**concat-0**), low level after the last convolutional layer (**concat-1**), medium level in the middle of the FC layers (**concat-2**), and high level just before the output (softmax) layer (**concat-3**).

Ideally, the streams should be fused when they have reached an optimal level of abstraction which remains to be seen empirically. In general, the optimal fusion level could depend on the task, data, discrepancies and importance of the information streams, and the architecture itself. However, the following could shed some lights on the benefits and liabilities of different schemes: first, assuming there is a fixed budget in terms of number of layers, placing the fusion point at the higher levels, leads to more layers being allocated to individual stream processing, leaving fewer layers and capacity for post-processing and abstraction extraction after fusion point. Second, the higher the fusion level, the higher the number of architecture parameters (*#params*). For example, *#params* of concat-3 is almost twice as many as concat-0.

5. EXPERIMENTAL RESULTS

5.1. Setup

DNNs were trained using PyTorch-Kaldi [25] with default recipes, without mono-phone regularisation. The architectures (Fig. 2) consists of a cascade of four 1D convolutional layers followed by five FC hidden layers. Experiments were carried out on TIMIT [26] and WSJ [10] tasks. Alignments were taken from the respective Kaldi standard recipes [27]. For TIMIT phone error rate (PER) and for WSJ word error rate (WER) is reported on the standard development (Dev) and evaluation (Eval) sets. Length of the MFCC, FBank and raw features (per frame) are 39, 80 and 257, respectively. For comparison purposes, the raw magnitude spectrum (Mag) and its 10^{th} root ($\text{Mag}^{0.1}$) were used, too. Feature vectors were augmented with the features of the ± 5 contextual frames.

5.2. Results and Discussion

Tables 1 and 2 show the PERs and WERs for TIMIT and WSJ databases, respectively. One interesting observation is that the wrapped phase spectrum (Fig. 1(c)) somehow returns surprisingly good results despite having a chaotic shape lacking any meaningful trend which could possibly facilitate the understanding or modelling process. This demonstrates the model is powerful enough to decipher such a complicated sequence.

On the other hand, phase unwrapping[1] worsens the performance. The main problem with it is *instability*. As shown in Fig. 1(d) even a tiny change in the signal via adding White noise in 40 dB using two different seeds, could significantly change the output. Moreover, the accumulative nature of unwrapping error further aggravates this issue. Using more advanced methods such as [28] may help towards getting more stable results and is recommended for future work.

In comparison with the wrapped and unwrapped phase spectra, using phase of the MinPh component (Phase-MinPh), computed via the Hilbert transform (Eq. (3)), leads to a higher performance. Besides, using its negative derivative, namely GD-MinPh, further lowers the PER/WER. In [20], it is argued that phase spectrum (approximately) behaves like a frequency modulated (FM) signal; differentiation demodulates it and turns the GD into an amplitude modulated (AM) signal. This could justify the higher performance of the GD, although further investigation is warranted to draw a firm conclusion.

The usefulness of the phase differentiation also highlights the importance of applying some pre-processing steps which may not affect the information content but could lead to a better information representation. Another example is compressing the dynamic range of the raw magnitude spectrum via root compression which results in a significant performance gain.

Acoustic modelling using the excitation part (GD-Exc), although admittedly unsuitable for ASR, leads to remarkably better PER/WER than the random choice. Considering this point, and the fact that it contains complementary information overlooked by the filter component, encourages simultaneous deployment of these two information streams using multi-head CNNs, as discussed in Section 4. As seen, source-filter fusion significantly improves the performance and helps the phase-based features to outperform the magnitude-based ones by a notable margin. Also note that multi-stream system with optimal fusion scheme, outperforms the sum of the GD-VT and GD-Exc, namely GD-MinPh.

The relative (to GD-VT) WER reduction after applying the source component via optimal fusion scheme is 11% and 6.7% for the WSJ (concat-1) and TIMIT (concat-2) respectively, which is a significant gain. The relative gain is higher for WSJ, owing to availability of more training data (81h vs 5.4h) which not only leads to a more effective training but also makes the WSJ results more reliable[2] for comparing different front-ends. As seen in Table 2, even without multi-stream processing, the raw phase-based features such as GD-MinPh or GD-VT outperform all of the magnitude-based ones.

[1] Unwrapping is done using *numpy.unwrap* command with default setting.

[2] We noticed TIMIT results could slightly vary across different runs.

Table 1. *TIMIT PER for different front-ends.*

	Dev	Eval
MFCC	17.1	18.6
FBank	16.3	18.2
Mag	16.8	17.8
$\text{Mag}^{0.1}$	15.9	17.6
Phase-Wrapped	21.6	23.7
Phase-UnWrapped	29.6	31.8
Phase-MinPh	16.8	18.6
GD-MinPh	16.9	18.4
GD-VT	18.2	19.3
GD-Exc	31.3	32.3
Concat-0	16.8	18.4
Concat-1	16.3	18.1
Concat-2	16.2	18.0
Concat-3	17.0	18.4

Table 2. *WSJ WER for different front-ends.*

	Dev-93	Eval-92	Eval-93
MFCC	10.4	6.8	10.4
FBank	9.1	5.9	8.8
Mag	9.3	5.9	9.1
$\text{Mag}^{0.1}$	**8.8**	**5.5**	**9.0**
Phase-Wrapped	9.9	6.1	10.4
Phase-UnWrapped	13.1	8.9	16.4
Phase-MinPh	9.3	5.8	9.4
GD-MinPh	**8.3**	**5.1**	**7.8**
GD-VT	8.6	5.4	7.6
GD-Exc	12.2	8.5	13.2
Concat-0	8.2	4.9	7.8
Concat-1	**7.9**	**4.8**	**7.4**
Concat-2	8.1	4.8	7.7
Concat-3	8.2	5.0	8.1

6. CONCLUSIONS

In this paper, we reviewed the phase spectrum applications in ASR and investigated the effectiveness of acoustic modelling from the raw phase spectrum. Acoustic models were successfully built from the raw wrapped, unwrapped, minimum-phase, excitation and vocal tract phase spectra through CNNs, leading to comparable to better results than the magnitude-based features on the TIMIT and WSJ tasks. Furthermore, we studied acoustic modelling using multi-head CNNs fed with the raw phase spectra of the source and filter components, resulting in up to 7.4% WER in WSJ (Eval-93) task. Employing the raw phase spectra of the source and filter components using multi-head CNNs for various speech recognition and classification tasks is a broad avenue for future research.

7. REFERENCES

[1] H. Helmholtz and A. Ellis, *On the sensations of tone as a physiological basis for the theory of music / by Herman L.F. Helmholtz*. Longmans, Green London, 1885.

[2] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[3] B. Yegnanarayana, "Formant extraction from linear prediction phase spectra," *JASA*, vol. 63, no. 5, pp. 1638 – 1640, 1978.

[4] F. Itakura and T. Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum," in *ICASSP*, vol. 12, 1987, pp. 1257–1260.

[5] H. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Processing*, vol. 22, no. 3, pp. 259 – 267, 1991.

[6] R. Hegde, H. Murthy, and V. Gadde, "Significance of the modified group delay feature in speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 190–202, Jan 2007.

[7] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Communication*, vol. 49, no. 3, pp. 159 – 176, 2007.

[8] E. Loweimi, S. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *ICASSP*, May 2013, pp. 7155–7159.

[9] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain." in *INTERSPEECH*. ISCA, 2015, pp. 598–602.

[10] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *ICASSP*, 1992, pp. 899–902.

[11] D. Pearce and H. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *INTERSPEECH*, 2000, pp. 29–32.

[12] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *ICASSP '01*, vol. 1, 2001, pp. 133–136 vol.1.

[13] Y. Wang, J. Hansen, G. Allu, and R. Kumaresan, "Average instantaneous frequency (AIF) and average log-envelopes (ALE) for ASR with the aurora 2 database," in *INTERSPEECH*. ISCA, 2003.

[14] D. Zhu and K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *ICASSP*, vol. 1, May 2004, pp. I–125–8 vol.1.

[15] E. Loweimi, J. Barker, and T. Hain, "Exploring the use of group delay for generalised vts based noise compensation," in *ICASSP*, 2018, pp. 4824–4828.

[16] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *ICASSP*, vol. 2. IEEE, 1996, pp. 733–736.

[17] E. Loweimi, J. Barker, and T. Hain, "Channel compensation in the generalised vector taylor series approach to robust asr," in *INTERSPEECH*, 2017, pp. 2466–2470.

[18] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal and Information Process, Mississippi State University, Tech. Rep., 2002.

[19] E. Loweimi, J. Barker, and T. Hain, "Statistical normalisation of phase-based feature representation for robust speech recognition," in *ICASSP*, 2017, pp. 5310–5314.

[20] E. Loweimi, J. Barker, O. Saz Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain," in *INTERSPEECH*, 2017, pp. 414–418.

[21] E. Loweimi, "Robust phase-based speech signal processing; from source-filter separation to model-based robust asr," Ph.D. dissertation, University of Sheffield, Sheffield, UK, Feb 2018.

[22] E. Loweimi, J. Barker, and T. Hain, "On the usefulness of the speech phase spectrum for pitch extraction," in *INTERSPEECH*. ISCA, 2018, pp. 696–700.

[23] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[24] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust asr," in *ICASSP*, 2019, pp. 6745–6749.

[25] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *IEEE ICASSP*, 2019.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[28] T. Drugman and Y. Stylianou, "Fast and accurate phase unwrapping," in *INTERSPEECH*, 2015, pp. 1171–1175.